Euphytica (2008) 159:259–273
DOI 10.1007/s10681-007-9485-4

Molecular and SNP characterization of two genome specific transcription factor genes *GhMyb8* and *GhMyb10* in cotton species

Chuan-Yu Hsu · Chuanfu An · Sukumar Saha · Din-Pow Ma · Johnie N. Jenkins · Brian Scheffler · David M. Stelly

Received: 7 February 2007 / Accepted: 14 June 2007 / Published online: 1 July 2007

Abstract Two *R2R3-Myb* cDNAs (*GhMyb8* and *GhMyb10*) and their corresponding genes were isolated and characterized from allotetraploid cotton (*Gossypium hirsutum* L. cv. DES119) fiber cells. Both *GhMyb8* and *GhMyb10* exhibit some conserved features shared in subgroup 4 of plant *R2R3-MYB* proteins, including the GIDxxH motif and a zinc-finger domain. Both genomic origin and single nucleotide polymorphism (SNP) analyses reveal that *GhMyb8* and *10* are alloallelic genes in the allotetraploid cotton (AD genome). *GhMyb10* is derived from the A_2 subgenome, whereas *GhMyb8* is from the D_5 subgenome. Possible chromosomal locations of these two genes were explored by SNP marker based deletion analyses. The results showed that the average rate of SNP per nucleotide among the selected genotypes in the two gene fragments was 3.75% (~one SNP per 27 nucleotide), and 0.55% and 5.14% in coding regions and 3′-UTR (3′ untranslated regions), respectively. Northern blot analysis showed that *GhMyb8* and *GhMyb10* are expressed in all examined tissues, including leaves, flowers, roots, and fibers from different developmental stages; however, the transcript level of *GhMyb8/10* is more abundant in flowers and roots. The ectopic expression of *GhMyb10* in transgenic tobacco plants showed the abnormal cell shapes in leaf trichomes, suggesting that *GhMyb8* and *GhMyb10* might play a role in the process of trichome cell differentiation. The exact chromosomal location of the two *myb* genes couldn't be determined using the SNP deletion method due to the incomplete coverage of cytogenetic stocks.

The nucleotide sequences of *GhMyb8* and *GhMyb10* have been submitted to GenBank and assigned accession numbers EF421795 and EF421796, respectively.

Electronic supplementary material The online version of this article (doi:10.1007/s10681-007-9485-4) contains supplementary material, which is available to authorized users.

C.-Y. Hsu · D.-P. Ma (✉)
Department of Biochemistry and Molecular Biology,
Mississippi State University, Mississippi State, MS
39762, USA
e-mail: dm1@ra.msstate.edu

C. An
Department of Plant and Soil Sciences, Mississippi State
University, Mississippi State, MS 39762, USA

S. Saha · J. N. Jenkins
USDA-ARS Crop Science Research Laboratory,
Mississippi State, MS 39762, USA

B. Scheffler
USDA-ARS, USDA-CGRU, Stoneville, MS 38776, USA

B. Scheffler
Department of Plant Sciences, University of California,
Davis, CA 95616, USA

D. M. Stelly
Department of Soil and Crop Sciences, Texas A&M
University, College Station, TX 77843, USA

Keywords *Gossypium hirsutum* · *R2R3-Myb* · Transgenic tobacco plants · Trichome

Introduction

MYB proteins, containing a conserved domain with DNA-binding ability (DBD), generally serve as transcriptional factors and regulate the transcriptional expression levels of downstream genes (Gonda 1998). In plants, a large number of *MYB* proteins, especially the subfamily of *R2R3-MYB* proteins which contain two imperfect repeats (R2 and R3) in the DBD region, are extensively expressed. This *R2R3-Myb* gene family is mainly involved in regulatory control of many plant-specific processes, including secondary metabolism, cell shape development, cell division, signal transduction, and disease resistance (Cone et al. 1993; Paz-Ares et al. 1987; Martin and Paz-Ares 1997).

Several plant *R2R3-Myb* genes are well known to mediate the differentiation process of epidermal cells. The *Arabidopsis GLABROUS 1* (*GL1*, *AtMybGL1*) has been identified as an essential gene for the initiation of leaf trichomes (Oppenheimer et al. 1991); whereas *MIXTA* from *Antirrhinum majus* has been shown to regulate the development of conical cells or multicellular trichomes in floral papillae (Noda et al. 1994; Glover et al. 1998). Since cotton fibers are single-celled seed trichomes differentiating from the epidermal cells of cotton ovules (Wilkins et al. 2000), *MYB* proteins have been proposed to function in regulation of differentiation and development of cotton fibers. Indeed, some *R2R3-Myb* genes have been recently isolated and characterized from cotton fibers (Loguercio et al. 1999; Suo et al. 2003; Wang et al. 2004; Hsu et al. 2005; Wu et al. 2006). Among them, *GhMyb1*, *2*, and *3* are abundantly expressed in all tissues; whereas *GhMyb4*, *5*, and *6* have lower, but tissue-specific expression patterns (Loguercio et al. 1999). Other *Myb* genes, such as *GhMYB109* and *GhMyb25*, exhibit fiber-specific expression patterns, with *GhMYB109* being induced specifically in fiber initials and elongated fiber cells (Suo et al. 2003), whereas *GhMyb25* is expressed only in fiber initials on the day of anthesis (Wu et al. 2006). Two alloallelic genes in allotetraploid cotton (*Gossypium hirsutum*, AD genome), *GhMyb7* and *9*, are expressed in flowers and fibers, with their expression in fibers being developmentally regulated (Hsu et al. 2005). In vitro DNA-binding assays have showed that *GhMYB7* protein plays a role in regulation of cotton fiber development by interacting with a fiber-specific promoter of *Ltp3* gene (Hsu et al. 2005). Similarly, *GaMYB2* isolated from *G. arboretum*, showed an interaction with a fiber-specific promoter of the *RDL1* gene in a yeast one-hybrid system (Wang et al. 2004). Thus, *MYB* proteins appear to play an important role in the differentiation and development of cotton fibers.

Single nucleotide polymorphism (SNP), including single DNA base differences plus small insertions and deletions, are the most abundant sequence variations in most genomes (Wang et al. 1998; Brookes 1999; Cho et al. 1999; Zimdahl et al. 2004). The abundance, ubiquity, and interspersed nature of SNP throughout the genome make them ideal candidates as molecular markers in characterizing allelic variation, QTL (quantitative trait loci) mapping, and marker-assisted selection in crops (Rafalski 2002). A number of reports have provided information about sequence diversity and SNP markers in *Arabidopsis thaliana* (L.) heynh, maize (*Zea mays* ssp. *mays* L.), rice (*Oryza sativa* L.), soybean (*Glycine max* L. Merr.), barley (*Hordeum vulgare* ssp. *spontaneum*), and wheat (*Triticum aestivum*) (Cho et al. 1999; Ching et al. 2002; Kanazin et al. 2002; Zhu et al. 2003; Caldwell et al. 2004; Feltus et al. 2004; Kim et al. 2005). In cotton (*Gossypium* spp.), the analysis of DNA sequence variation has focused primarily on single genes or DNA fragments, with the aim of defining evolutionary relationships of species (Cronn et al. 2002b; Senchina et al. 2003). Most DNA markers are separated based on length differences between alleles at a locus. The alleles of many genes of interest, however, have the same length but contain DNA sequence difference. SNP markers derived from functional genes can be used as a tool for candidate gene mapping and provide valuable information in molecular mapping of QTLs.

In this study, two alloallelic *R2R3-Myb* genes (*GhMyb8* and *GhMyb10*) were isolated from a fiber cDNA library of allotetraploid (AD genome) cotton. The expression of *GhMyb8/10* transcript and its possible physiological role were characterized. The sequence variations of these two alloallelic *Myb* genes among the selected cotton species were determined and SNP markers were developed for chromosomal assignment.

Materials and methods

Isolation of *GhMyb8* and *GhMyb10* genes

Cotton (*Gossypium hirsutum* L. cv. DES119) plants were grown in a greenhouse at USDA/ARS, Mississippi State or annually planted in the field at the North Farm at Mississippi State University. Cotton flowers were tagged on the day of anthesis (0 DPA), and fibers were collected at different developmental stages (5, 10, 15, and 20 DPA).

Total RNA was isolated from 15 DPA fibers using a modified method of Hughes and Galau (1988). The construction of an adaptor-ligated double-strand fiber cDNA library was conducted using a Marathon cDNA Amplification Kit (BD Biosciences, Clontech, Palo Alto, CA) according to the manufacturer's instructions. The conserved region of *R2R3-Myb* genes was amplified by PCR using cotton fiber cDNA library as a template and a pair of degenerated primers (*Myb-deg-F* and *Myb-deg-R*, see supplementary Table 1). The PCR product (190 bp) was cloned into pGEM-T easy vector (Promega, Madison, WI), and the resulting recombinant plasmids were analyzed and sequenced using an ABI PRISM 310 DNA Genetic Analyzer (Perkin-Elmer, Applied Biosystems, Foster City, CA). Based on the sequence of the 190 bp PCR fragment, a pair of gene-specific primers (*Myb-5* and *Myb-6*, see supplementary Table 1) were designed and used in Rapid Amplification of cDNA Ends (RACEs) with the Marathon cDNA Amplification kit (Clontech). In the 5′-RACE reaction, the gene-specific primer (*Myb-6*) and the Adaptor primer AP1 (supplementary Table 1) were used in the first PCR, and the Adaptor primer AP2 (supplementary Table 1) was then used in the nested PCR. The first and nested 3′-RACEs were performed as the 5′-RACE but using the gene-specific primer *Myb-5*. Two full length cDNAs (*GhMyb8* and *GhMyb10*) were amplified, cloned, and sequenced.

The 5′- and 3′-flanking regions of *GhMyb8* and *10* genes were cloned using a PCR-based genomic walking method (Siebert et al. 1995). Cotton genomic DNA (3 μg) was digested with *Sca*I, ligated with the Marathon adaptor (Clontech), and then used as a template in the PCR amplifications of genomic walking. The amplifications of 5′- and 3′-genomic walking were conducted as similarly to 5′- and 3′-RACEs, except using *Myb-10* and *Myb-9* primers (supplementary Table 1) for first PCR amplifications, respectively. Moreover, the *Myb-6* and *Myb-5* primers were used in the nested PCR amplifications of 5′- and 3′-genomic walking, respectively. The nested PCR products were purified, cloned and sequenced as previously described. The final PCR amplifications with *Pfu* DNA polymerase (Stratagene, La Jolla, CA), containing full-length *GhMyb8* and *GhMyb10* cDNAs and their corresponding genes, were cloned and sequenced. At least two individual clones from each recombinant construct were sequenced to further confirm sequence accuracy of *GhMyb8* and *GhMyb10* genes.

Northern and genomic Southern analyses

Northern analysis was carried out using total RNA (10 μg) isolated from different cotton tissues, including leaves, flowers, roots, and fibers at different developmental stages (5, 10, 15, and 20 DPA). After electrophoresis, RNA samples were blotted and hybridized with the C-terminal region (transregulatory region, TRR) of *GhMyb10* cDNA labeled by [α-^{32}P] dCTP with the random priming method (Feinberg and Vogelstein 1983). The hybridization signal on the membrane was then detected by autoradiography. For Southern blotting, genomic DNA (10 μg) isolated from cotton leaves using a modified method of Paterson et al. (1993) was individually digested with six restriction enzymes, *Dra*I, *Eco*RI, *Eco*RV, *Sca*I, *Ssp*I, or *Xba*I. The digested DNAs were separated by electrophoresis, blotted, and then hybridized by using a similar protocol as Northern analysis.

Genomic origin and SNP analyses of *GhMyb8* and *GhMyb10* genes

Five different *Gossypium* species were used in genomic origin analysis, including *G. herbaceum* (diploid A_1 genome, accession number A_1-57), *G. arboreum* (diploid A_2 genome, accession number A_2-86), *G. thurberi* (diploid D_1 genome, accession number D_1-1), *G. raimondii* (diploid D_5 genome, accession number D_5-4), and *G. hirsutum* L. cv. DES119 (allotetraploid AADD genome).

In SNP analyses, eight genotypes from six species of cotton (*Gossypium* spp.) were used as plant materials, including two diploid species (*G. arboreum* L. [A_2] and *G. raimondii* Ulbrich [D_5]) and four tetraploid species: TM-1, HS46, MARCABUCAG8US-1-88 (*G. hirsutum* L., AD_1), 3-79 (*G. barbadense* L., AD_2), *G. tomentosum* Nuttall ex Seemann (AD_3), and *G. mustelinum* Miers ex Watt (AD_4). TM-1 and 3-79 were the genetic standards for cultivated *G. hirsutum* and *G. barbadense*, respectively. HS46 and MARCABUCAG8US-1-88 are two parental lines used for developing recombinant inbreed lines (Shappley et al. 1998a, 1998b).

Three kinds of genetic stock were used for chromosomal assignment of the SNP markers by deletion analysis method (Liu et al. 2000; Ulloa et al. 2005). The stocks included: (1) interspecific F_1 hybrid hypoaneuploid chromosome substitution stock composed quasi-isogenic monosomic (2n = 51) and arm deficient monotelodisomic (2n = 52) F_1 interspecific hybrids between the Upland cotton (*G. hirsutum*) inbred TM-1 and one of two species, either *G. barbadense* 3-79 or *G. tomentosum*. Monotelodisomes included chromosomes 1Sh, 1Lo, 2sh, 2Lo, 3sh, 3Lo, 4sh, 4Lo, 5Lo, 6sh, 6Lo, 7sh, 7Lo, 8Lo, 9Lo, 10sh, 10Lo, 11Lo, 12Lo, 14Lo, 15Lo, 16sh, 16Lo, 17sh, 18sh, 18Lo, 20sh, 20Lo, 22sh, 22Lo, 25Lo, 26sh, and 26Lo. Monosomes included chromosomes 1, 2, 4, 6, 7, 9, 10, 12, 16, 17, 18, 20, 23, and 25. The monosomic stocks were labeled for the missing *G. hirsutum* (TM-1) chromosome. The monotelodisomic stocks were labeled by the particular chromosome arm that is present; (2) monosomic reciprocal translocation lines in *G. tomentosum* (NTN lines), which consisted of NTN10-19, NTN17-11, NTN4-15, NTN6-14, NTN12-11, NTN16-15, and NTN7-11. The reciprocal translocation lines are duplicate-deficient stocks missing a chromosome segment of two chromosomes involved in reciprocal translocation. For example, NTN 7-11 sub F_1 denotes that *G. hirsutum* chromosome lacks chromosome segments of chromosomes 7 and 11; (3) euploid interspecific chromosome substitution lines (CS-B, BC_5S_1) of *G. barbadense* in TM-1, included 01, 02, 04, 5Sh, 06, 07, 09, 10, 11Sh, 12, 12Sh, 14Sh, 15Sh, 16, 17, 18, 22Sh, 22Lo, 25, and 26Lo, which named as specific 3-79 chromosome or chromosome part in TM-1 background.

Genomic DNAs of the plant materials were isolated with a Qiagen DNeasy plant maxi kit (Qiagen Inc, Valencia, CA) following the manufacturer's protocol. DNA samples of the two diploid species, *G. arboreum* (A_2) and *G. raimondii* Ulbrich (D_5), were kindly provided by Dr. John Yu (USDA/ARS, Crop Germplasm Research Unit, College Station, TX).

The gene-specific primer pairs, *GhMyb8-F*/*GhMyb8-g2* and *GhMyb10-F*/*GhMyb10-g2* (supplementary Table 1), were used for the amplification of *GhMyb8* and *GhMyb10* genes in both genomic origin and SNP analyses. In SNP analysis, the PCR products amplified by *Pfu* polymerase (Stratagene, La Jolla, CA) were cloned into TOPO TA vector (Invitrogen, Carlsbad, CA) and twelve individual colonies from each recombinant construct were sequenced using the ABI PRISM 3730XL DNA Genetic Analyzer (Applied Biosystems, Foster City, CA). For each gene fragment, a minimum of forward and reverse matched sequences from three clones were considered as a corrected sequence and used to determine the possible duplicated copy of each gene (Cronn et al. 2002a; Cedroni et al. 2003; Rong et al. 2004).

The sequence alignment was conducted by DNASTAR software (DNASTAR Inc., Madison, Wisconsin, USA), and DnaSP 4.0 software (Rozas et al. 2003) was used for SNP identification and character analysis. Seven interspecies SNP primers (TM-1 and 3-79 or TM-1 and *G. tomentosum*, supplementary Table 1) were designed from just the upstream or downstream of the SNP site in different regions, so that SNP markers could be detected by single base extension technique.

ABI Prism SNaPshot^{TM} multiplex kits with an ABI 3100 system were used for SNP genotyping with a slightly modified protocol in this experiment (Applied Biosystems, Foster City, CA). The *Pfu* DNA polymerase amplified PCR products from the genomic DNA of genetic stocks were purified by enzyme SAP (shrimp alkaline phosphatase) and *Exo* I (2 units of SAP and 4 units of *Exo* I for 20 μl PCR product) at 37°C for 1 h followed by 75°C for 15 min. Single base extension was performed in a 7 μl reaction mixture containing 1.5 μl of SnaPshot Multiplex Ready Reaction Mix, 0.5 μl of purified PCR product, 0.2 μl of SNP primer (10 μM), and 4.3 μl of distilled water. The thermal cycling parameters included 25 cycles of 96°C, 10 s, 50°C, 5 s, and 60°C, 30 s. After treating with 1 unit SAP at 37°C for 1 h followed by 75°C for 15 min, 1 μl of 10x diluted SnaPshot product was analyzed with the ABI 3100 Genetic Analyzer system.

Overexpressing *GhMyb10* in transgenic tobacco plants

The full-length coding region of *GhMyb10* was amplified by PCR using *Pfu* DNA polymerase (Stratagene, La Jolla, CA) and *GhMyb10-E3* and *GhMyb10-E4* primers (supplementary Table 1), and cloned into a binary vector pBI121 (Clontech, Palo Alto, CA) downstream of the CaMV 35S promoter without the GUS reporter gene. The recombinant plasmids were transferred into *Agrobacterium tumefaciens* LBA4404 cells by a freeze-thaw method (Walkerpeach and Velten 1994). The transgenic tobacco plants (*Nicotiana tabacum* L.) were generated with transformed *A. tumefaciens* LBA4404 cells using the standard leaf-disk transformation method (Gallagher 1992; Jefferson et al. 1987). The morphology of leaf trichomes were examined and photographed under a stereo microscope (Olympus, model SZH10 with Olympus Camera System, Center Valley, PA).

Results

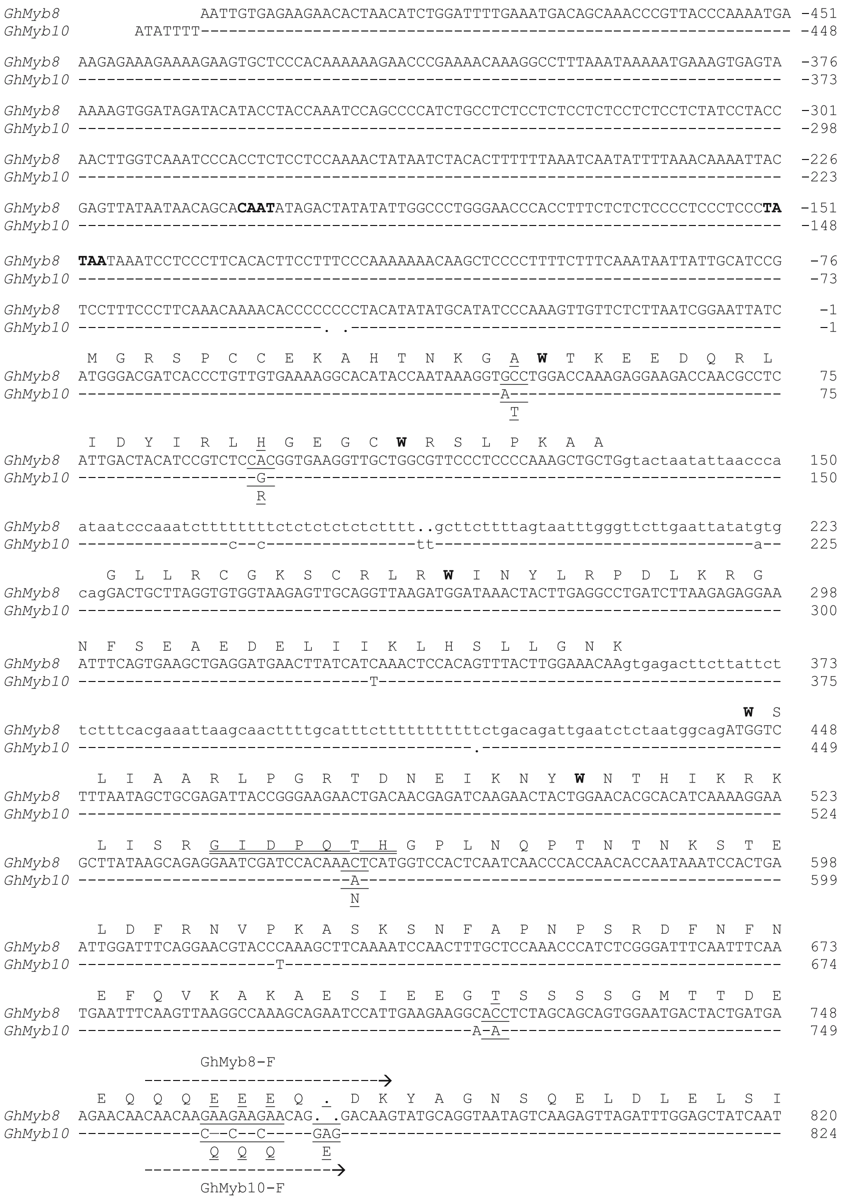
Cloning and characterization of *GhMyb8* and *GhMyb10* cDNAs and their corresponding genes

Two full-length cDNAs, named as *GhMyb8* and *GhMyb10* (Gh represents *G. hirsutum*), encoding *R2R3-MYB* proteins, and their corresponding genomic DNA fragments (2.14 kb and 2.18 kb, respectively) were isolated from cotton. The determined nucleotide (nt) and deduced amino acid (aa) sequences of *GhMyb8* and *GhMyb10* along with their 5′- and 3′-flanking regions are shown in Fig. 1. *GhMyb8* encodes a protein containing 284-aa, whereas *GhMyb10* encodes a 281-aa protein. Both *GhMYB8* and *GhMYB10* proteins exhibit common characteristics of plant *R2R3-MYBs*: a conserved DNA-binding domain (DBD) containing two (R2 and R3) imperfect repeats at the N-terminus with three regularly spaced tryptophan (W) residues and a nonconserved transcriptional regulatory region (TRR) at the C-terminus of proteins. Furthermore, both *GhMYB8* and *GhMYB10* proteins also contain some conserved features found in the subgroup 4 of *Arabidopsis R2R3-Myb* gene family (Kranz et al. 1998; Stracke et al. 2001), including a basic TRR1 region (40-aa immediately downstream of the DNA-binding domain), a GIDPxxH motif (x represents any aa residue), and a cysteine-rich zinc-finger domain. Since both *GhMYB8* and *GhMYB10* proteins contain an overall acidic TRR region at the C-termini of proteins, it is believed that this region may serve as a transcriptional activator.

In comparison with the R2R3-regions of other plant *R2R3-MYB* proteins, *GhMYB8* and 10 share about 95.7% and 93.9% amino acid identities, respectively, with *GhMYB1* (Loguercio et al. 1999), and about 92.2% and 90.4% identities, respectively, with *AtMYB4* (Kranz et al. 1998), suggesting that these four MYB proteins may recognize similar DNA-binding motifs. A phylogenetic tree (Fig. 2) constructed by comparing the deduced amino acid sequences of full-length *GhMYB8* and 22 *R2R3-MYBs* from different plant species indicates that *GhMYB8* and *GhMYB10* belong to the subgroup 4 of the *A. thaliana R2R3-MYB* family (Kranz et al. 1998; Stracke et al. 2001) as *GhMYB1* and *AtMYB4*. The *GhMyb8* and *GhMyb10* genes (Fig. 1) contain two introns located at their N-terminal DBD regions. The sizes of the corresponding introns between *GhMyb8* and *10* are similar, and the positions of two individual introns are also conserved. Overall, *GhMYB8* and *GhMYB10* proteins share high amino acid identities in both *R2R3*-conserved region (about 98.3%) and full-length (including nonconserved TRR region) protein (about 94%). Their corresponding genes have slightly less identity (82.1%) in the nucleotide sequences. These high levels of identities in both amino acid and nucleotide sequences suggest that *GhMyb8* and *GhMyb10* genes in the allotetraploid cotton (*G. hirsutum* L. cv. DES119) genome (AADD genome) are alloallelic genes and derived from different genomic origins (see below).

Southern blot and genomic origin analyses of *GhMyb8* and *GhMyb10* genes

The allotetraploid cotton (*G. hirsutum* L. cv. DES119) genomic DNA was digested with six restriction enzymes, *Dra*I, *Eco*RI, *Eco*RV, *Sca*I, *Ssp*I,



```
GhMyb8                 AATTGTGAGAAGAACACTAACATCTGGATTTTGAAATGACAGCAAACCCGTTACCCAAAATGA -451
GhMyb10         ATATTTT--------------------------------------------------------------- -448

GhMyb8    AAGAGAAAGAAAAGAAGTGCTCCCACAAAAAAGAACCCGAAAACAAAGGCCTTTAAATAAAAATGAAAGTGAGTA  -376
GhMyb10   ---------------------------------------------------------------------------  -373

GhMyb8    AAAAGTGGATAGATACATACCTACCAAATCCAGCCCCATCTGCCTCTCCTCTCCTCTCCTCTCCTCTATCCTACC  -301
GhMyb10   ---------------------------------------------------------------------------  -298

GhMyb8    AACTTGGTCAAATCCCACCTCTCCTCCAAAACTATAATCTACACTTTTTTAAATCAATATTTTAAACAAAATTAC  -226
GhMyb10   ---------------------------------------------------------------------------  -223

GhMyb8    GAGTTATAATAACAGCACAATATAGACTATATATTGGCCCTGGGAACCCACCTTTCTCTCTCCCCTCCCTCCCTA  -151
GhMyb10   ---------------------------------------------------------------------------  -148

GhMyb8    TAATAAATCCTCCCTTCACACTTCCTTTCCCAAAAAAACAAGCTCCCCTTTTCTTTCAAATAATTATTGCATCCG   -76
GhMyb10   ---------------------------------------------------------------------------   -73

GhMyb8    TCCTTTCCCTTCAAACAAAACACCCCCCCCTACATATATGCATATCCCAAAGTTGTTCTCTTAATCGGAATTATC    -1
GhMyb10   --------------------------. .----------------------------------------------    -1

           M  G  R  S  P  C  C  E  K  A  H  T  N  K  G  A  W  T  K  E  E  D  Q  R  L
GhMyb8    ATGGGACGATCACCCTGTTGTGAAAAGGCACATACCAATAAAGGTGCCTGGACCAAAGAGGAAGACCAACGCCTC    75
GhMyb10   ---------------------------------------------A-----------------------------    75
                                                        T

           I  D  Y  I  R  L  H  G  E  G  C  W  R  S  L  P  K  A  A
GhMyb8    ATTGACTACATCCGTCTCCACGGTGAAGGTTGCTGGCGTTCCCTCCCCAAAGCTGCTGgtactaatattaaccca   150
GhMyb10   -------------------G-------------------------------------------------------   150
                             R

GhMyb8    ataatcccaaatctttttttctctctctctcttttt..gcttcttttagtaatttgggttcttgaattatatgtg   223
GhMyb10   ----------------c--c----------------tt----------------------------------a--   225

              G  L  L  R  C  G  K  S  C  R  L  R  W  I  N  Y  L  R  P  D  L  K  R  G
GhMyb8    cagGACTGCTTAGGTGTGGTAAGAGTTGCAGGTTAAGATGGATAAACTACTTGAGGCCTGATCTTAAGAGAGGAA   298
GhMyb10   ---------------------------------------------------------------------------   300

          N  F  S  E  A  E  D  E  L  I  I  K  L  H  S  L  L  G  N  K
GhMyb8    ATTTCAGTGAAGCTGAGGATGAACTTATCATCAAACTCCACAGTTTACTTGGAAACAAgtgagacttcttattct   373
GhMyb10   -------------------------------T-------------------------------------------   375

                                                                                 W  S
GhMyb8    tctttcacgaaattaagcaacttttgcatttctttttttttttctgacagattgaatctctaatggcagATGGTC   448
GhMyb10   ------------------------------------------.--------------------------------   449

            L  I  A  A  R  L  P  G  R  T  D  N  E  I  K  N  Y  W  N  T  H  I  K  R  K
GhMyb8    TTTAATAGCTGCGAGATTACCGGGAAGAACTGACAACGAGATCAAGAACTACTGGAACACGCACATCAAAAGGAA   523
GhMyb10   ---------------------------------------------------------------------------   524

            L  I  S  R  G  I  D  P  Q  T  H  G  P  L  N  Q  P  T  N  T  N  K  S  T  E
GhMyb8    GCTTATAAGCAGAGGAATCGATCCACAAACTCATGGTCCACTCAATCAACCCACCAACACCAATAAATCCACTGA   598
GhMyb10   -----------------------------A---------------------------------------------   599
                                       N

            L  D  F  R  N  V  P  K  A  S  K  S  N  F  A  P  N  P  S  R  D  F  N  F  N
GhMyb8    ATTGGATTTCAGGAACGTACCCAAAGCTTCAAAATCCAACTTTGCTCCAAACCCATCTCGGGATTTCAATTTCAA   673
GhMyb10   ---------------------T-----------------------------------------------------   674

            E  F  Q  V  K  A  K  A  E  S  I  E  E  G  T  S  S  S  S  G  M  T  T  D  E
GhMyb8    TGAATTTCAAGTTAAGGCCAAAGCAGAATCCATTGAAGAAGGCACCTCTAGCAGCAGTGGAATGACTACTGATGA   748
GhMyb10   ------------------------------------------A-A------------------------------   749

                       GhMyb8-F
                 -------------------------->
            E  Q  Q  Q  E  E  E  Q  .  D  K  Y  A  G  N  S  Q  E  L  D  L  E  L  S  I
GhMyb8    AGAACAACAACAAGAAGAAGAACAG. .GACAAGTATGCAGGTAATAGTCAAGAGTTAGATTTGGAGCTATCAAT   820
GhMyb10   -------------C--C--C-----GAG-----------------------------------------------   824
                        Q  Q  Q     E
                 --------------------->
                       GhMyb10-F
```

Fig. 1 Comparison of nucleotide and derived amino acid Sequences of *GhMyb8* and *GhMyb10* genes. The basal promoter elements (TATA and CAAT boxes) and the polyadenylation signal (AAATATA) are in bold. Introns are in lowercase letters. The translational stop codon is marked with an asterisk. The first nt (A) of the translation start codon (ATG) is assigned as position 1 in the nt sequence, and the nt positions upstream of position 1 are presented with minus numbers. The identical nucleotides are represented with dash lines, and the gaps are marked as dot lines. Differences in amino acid sequences between *GhMYB8* and *GhMYB10* proteins are underlined. The conserved R2 and R3 imperfect repeat regions are located at aa 12–64 and aa 65–115, and the regularly spaced tryptophan (W) residues of each repeat are in bold. Other conserved features in some plant *R2R3-MYB* proteins, including the GIDPxxH motif and the zinc-finger domain are double-underlined. Four PCR primers, *GhMyb8-F*, *GhMyb8-g2*, *GhMyb10-F* and *GhMyb10-g2* used in the determination of genomic origins are labeled

and *Xba*I, and hybridized with the *GhMyb10*-spcific probe (TRR region). Southern genomic blotting (Fig. 3a) showed one, two or three hybridized DNA fragments, indicating that the *GhMyb10* gene contained at least two similar copies in the allotetraploid cotton genome.

Fig. 1 continued

```
            G  I  S  S  S  G  K  N  N  N  S  T  G  V  S  T  A  N  S  A  E  S  K  P  L
GhMyb8    TGGGATTAGTTCATCCGGAAAGAACAACAACTCAACTGGGGTTTCAACTGCTAACTCAGCCGAATCCAAACCGCT   895
GhMyb10   ------------------------------------C--------------------------------------   899

            L  D  K  S  N  F  Q  F  L  G  Q  A  M  A  A  K  A  V  C  L  C  C  Q  L  G
GhMyb8    GTTAGACAAAAGCAATTTCCAGTTTCTTGGACAAGCTATGGCGGCTAAAGCAGTCTGTTTGTGTTGCCAGTTAGG   970
GhMyb10   ---------------------------------G--------------------------------G--------   974
                                                                           W

            F  G  T  S  E  I  C  R  N  C  Q   S  T  N  G  F  N  T  Y  C  *
GhMyb8    GTTCGGAACAAGTGAAATTTGCAGGAACTGTCAA.AGTACAAATGGGTTTAATACATATTGTTGACCTTGGGATT  1044
GhMyb10   ----------------------------------A----------------------------------------  1049
                                             K  Y  K  W  V  *

GhMyb8    CATATAGTGCTCAATATTATTCTATTTTTCTTGTTTTGAGAAAATGGTGGACATAAATGCTTAATTTACTAACCC  1119
GhMyb10   -----------------G--------------------------A---A-----G--------------------  1124

GhMyb8    AAGCTAAATTACAAACACAAGGTGTTTGCTATGCTTTATTATTGAA. . . . . . . .AGAAATTACAAATT  1179
GhMyb10   ---------------------A------------------------TTTGTTGAAGGAAAA--------------  1199

GhMyb8    ACACTATATCAAATTTG. . . . . . . . . . .CTCCCATAACATTTAAAAATTTCCTTGAATTTTATCA  1233
GhMyb10   ----------C------ATAATAATAATAATAATAAAG-----------------------------------A-  1274

GhMyb8    AAA.CAATATTAATATAAATTTCCGTTTCTAATAAG.AAATACAATTAAAATACGTAACCCAATCAATTGATGTA  1306
GhMyb10   ---AT-A--------------G------T---A-AT-----------------------TA------C-------  1349

GhMyb8    TCATGGTTTTATTATCTTTTTACTATAAAACTAGACTACGGATAAATTTTCATTTTTATCAAAAAAATATAATTT  1381
GhMyb10   -A------------------.------TGGT--C----T-----------T--------.----------. . .  1418

GhMyb8    TACTACTAAATTATTAGAAAGATTTTATTTAAGTTACTCAATTATTCAAAAGTTTT.ATTTAAGTTAAAAAAATT  1455
GhMyb10   . . . . . . . . . . -T-----------C-G----------T---------T-------------G-T--  1472

GhMyb8    TCTCAACGAGCTTCAAGTGACAACTTGACGATTAATATGGTGGATTAGTACCCATTAACGAATAAAAAATTATAC  1530
GhMyb10   -GCT-T-----------C--TG-T--A------G---------------------C--------------C----  1547

GhMyb8    TCTACATCCAAATCATTCA. . .TTA..GAGGCAAAAGCTC. . TTCCATGAAAGGAAACTGAACTATAAAAG  1594
GhMyb10   -T-G---------------GACAA-C-AT---ATTTG--A-AAAAG--TT------A------------------  1622

                            GhMyb8-g2
                     ←-------------------------
GhMyb8    AAAAGGAGAAAGAGAGCTCTTAATTGATACAGACAGT. .                                     1631
GhMyb10   -. .--------------------G---GT--G--A-GCG                                     1659
                       ←--------------------------
                              GhMyb10-g2
```

Upland cotton (*Gossypium hirsutum* L.), one of predominant cultivated cottons, is an allotetraploid species with an AD genome and derived from an ancient cross between Old World (diploid species of A genome) and New World (diploid species of D genome) cottons (Percival and Kohel 1990). Genomic DNAs isolated from five cotton species with different genomes were used to analyze the genomic origins of *GhMyb8* and *GhMyb10* genes (Fig. 3b). A 876-bp *GhMyb8* fragment was amplified with the genomice DNAs isolated from *G. raimondii* (diploid D_5 genome) and *G. hirsutum* L. cv. DES119, whereas a 903-bp *GhMyb10* fragment was shown in *G. arboreum* and *G. hirsutum* L. cv. DES119 DNA templates. The results indicate that *GhMyb8* and *GhMyb10* are alloallelic genes in the allotetraploid cotton genome (AADD), with *GhMyb8* derived from the D subgenome of allotetraploid AADD genome and *GhMyb10* inherited from the A genome ancestor in the AADD genome cotton. These results also suggest that *G. arboreum* (diploid A_2 genome) and *G. raimondii* (diploid D_5 genome) might be the genome donors of the cultivated cotton, *G. hirsutum* L. cv. DES119 (allotetraploid AADD genome).

Analysis of SNP in genome-specific *GhMyb8* and *GhMyb10* genes

The 12 colonies of PCR-amplified DNA fragments of *GhMyb8* and *GhMyb10* genes having the same sequence suggested the absence of any heterozygous allele and orthologs in the tested genotypes. Due to the genome specific character of these two genes (*GhMyb8* in D genome and *GhMyb10* in A genome, Fig. 3b), seven sequences generated from the eight genotypes for SNP assay, as well as the original nucleotide sequence from *G. hirsutum* L. cv. DES119 (AD) were used for SNP analysis. The aligned lengths of DNA fragments in *GhMyb8* and *GhMyb10* were 876 bp and 911 bp, respectively.

A total of 23 SNP including one indel were identified in the *GhMyb8* gene fragment. Among them, only one G/C transversion SNP and one A indel were found in the 276-bp coding region and 600-bp of 3′-UTR, respectively. The partial *GhMyb10* gene fragment had a coding region of 267 bp and a 3′-UTR region of 644 bp, with a total of 44 SNP including one G/T transversion and one A/G transition SNP in the coding region. Results also showed that 31 of total 44 SNPs (~70.5%) were

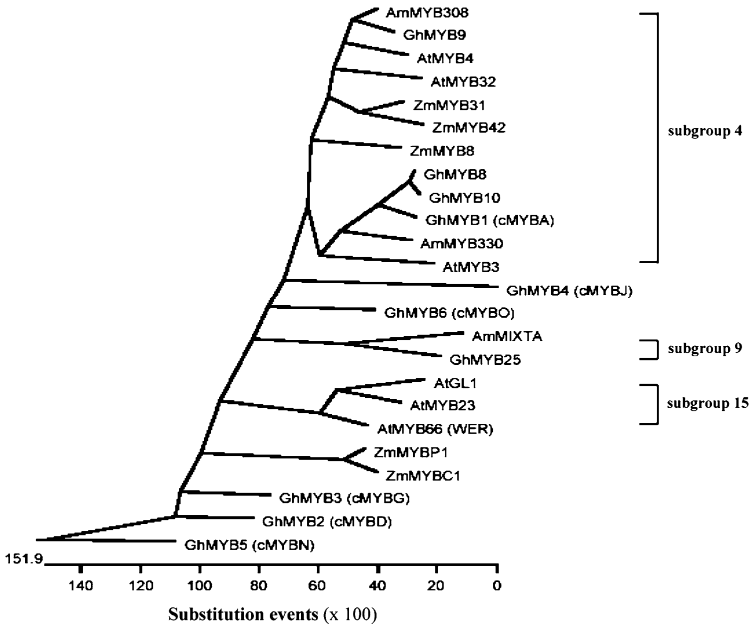


Fig. 2 A phylogenetic tree constructed by comparing amino acid sequences of *R2R3-MYB* proteins from four different plant species. The phylogenetic tree was constructed by the Clustal W method using the MegAlign program (DNASTAR Inc.) based on the amino acid sequences of *GhMYB8*, *10*, and twenty-two other plant *R2R3-MYB* proteins, including *AtGL1* (Oppenheimer et al. 1991), *AtMYB66* (WER) (Lee and Schiefelbein 1999), *AtMYB3* (Kranz et al. 1998), *AtMYB4* (Kranz et al. 1998), *AtMYB23* (Kitik et al. 2001) and *AtMYB32* (NP_195225) from *Arabidopsis thaliana*, *ZmMYBC1* (Paz-Ares et al. 1987), *ZmMYBP1* (Cone et al. 1993), *ZmMYB8* (Fornale et al. 2006), *ZmMYB31* (Fornale et al. 2006), and *ZmMYB42* (Fornale et al. 2006) from maize (*Zea mays*), *AmMIXTA* (Noda et al. 1994; Glover et al. 1998), *AmMYB308* (Tamagnone et al. 1998a), and *AmMYB330* (Tamagnone et al. 1998a) from snapdragon (*Antirrhinum majus*), and *GhMYB1* (Loguercio et al. 1999), *GhMYB2* (Loguercio et al. 1999), *GhMYB3* (Loguercio et al. 1999), *GhMYB4* (Loguercio et al. 1999), *GhMYB5* (Loguercio et al. 1999), *GhMYB6* (Loguercio et al. 1999), *GhMYB9* (AAK19619) *GhMYB25* (Wu et al. 2006) from *Gossypium hirsutum*. Subgrouping of twenty-four *R2R3-MYB* proteins was according to Kranz et al (1998)

indels, which were located in the 3′-UTR of *GhMy10*. One triallelic SNP site (A/T/C) was discovered at position 1511 of *GhMyb8*, as also found in soybean (Van et al. 2005). Theoretically, each SNP marker can have up to four possible alleles (A, C, G, and T), however, normally, only two alleles usually are present at any given SNP (e.g., C or T). This is possibly due to the low rate of mutation or base substitution at the nucleotide level. A polymorphic site of ‘TAA’ repeat motif was identified in the 3′-UTR of gene *GhMyb10* as demonstrated by Kumar et al. (2006).

The average rate of SNP per nucleotide in the two gene fragments was 3.75% (~one SNP/27bp nucleotide), and with 0.55% and 5.14% occurring in coding regions and 3′-UTRs, respectively. There was bias toward A and T nucleotides without any presence of C in the total 32 indels found in both gene fragments, similar to the report in maize (Batley et al. 2003). Of the 35 single-base changes in both gene fragments, transitions accounted for 21 (60%) and transversions for 13 (37%). In the total 4,344 bp of coding sequences analyzed, two of the three cSNP (SNP site in the gene coding region) were detected at the third codon position and the remaining one cSNP was at the second position of codon. All three cSNP were at the interspecific level (Tables 1 and 2).

The detailed reports on sequence variations among the eight genotypes for each of the two gene fragments are summarized in Tables 1 and 2. Only six indels had been identified at the intraspecific level of *G. hirsutum* in the 3′-UTR of *GhMyb10* gene. The other 61 variation sites (91%) were detected at the interspecific level. The result suggested the conserved character of *GhMyb8* and *GhMyb10* genes in *G. hirsutum*. In *GhMyb8*, the most polymorphic sequences were between *G. ramondii* and 3-79 (*G. barbadense*) (Table 1), whereas in *GhMyb10*,

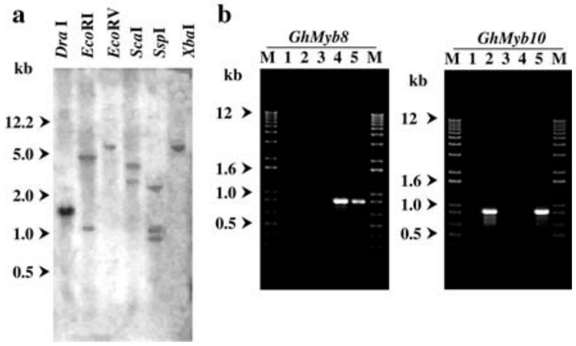


Fig. 3 Analysis of genomic composition and genomic origins of *GhMyb8* and *GhMyb10* genes. Genomic composition of *GhMyb8/10* gene was analyzed by Southern blotting (**a**) using DES119 genomic DNA (10 μg) digested with *Dra*I, *Eco*RI, *Eco*RV, *Sca*I, *Ssp*I, and *Xba*I and hybridized with the TRR region of *GhMy10*. The genomic origins were analyzed by PCR amplification (**b**) using *GhMyb8*- and *GhMyb10*- gene-specific primers on genomic DNAs from five different cotton species, *G. herbaceum* (diploid A_1 genome) (lane 1), *G. arboreum* (diploid A_2 genome) (lane 2), *G. thurberi* (diploid D_1 genome) (lane 3), *G. raimondii* (diploid D_5 genome) (lane 4), and *G. hirsutum* L. cv. DES119 (allotetraploid AADD genome) (lane 5). Lane M represents 1 kb plus DNA ladder

the largest variation existed between *G. arboreum* and *G. mustelinum* (Table 2). Based on their similarity with the diploid ancestral species, the results suggested two different origins of the haplotypes for *GhMyb8* and *GhMyb10* genes in the tetraploid cotton (supplementary Tables 2 and 3). The haplotypes also could broadly be separated into putative A genome and putative D genome group for the respective *GhMyb8* and *GhMyb10* gene based on their similarity with *G. arboreum* (A_2) and *G. raimondii* (D_5) species. Phylogenetic analyses with PAUP and other methods (Swofford 2003) suggested that independent evolution occurred in the two genomes of A and D in the tetraploid species after the events of polyploidy (data not shown). Similar results were observed in other cotton genes (Cronn et al. 1999).

Identification of possible chromosomal location of *GhMyb8* and *GhMyb10* genes

Seven interspecies SNP markers between TM-1 and 3-79 or TM-1 and *G. tomentosum* were targeted to identify the chromosomal location of these two genes using a deletion analysis method (supplementary Table 1). We did not find any missing SNP markers in any of the aneuploid cytogenetic stocks used in this study in the deletion analyses, suggesting that the two genes were not located to any of the missing chromosome of the respective cytogenetic stock. The failure to identify the exact chromosomal locations of these two genes is due to the incomplete coverage of cytogenetic stocks. The two *GhMyb* genes therefore could be located on the chromosomes or chromosome arm for which cytogenetic stocks are not available. Based on the subgenome specific character of the two genes and the results of deletion analyses, *GhMyb8* could be located on one of the

Table 1 SNP characters based on the *GhMyb8* gene fragment[a]

Genotype[b]	*G. ramondii*	3-79	*G. tomentosum*	*G. mustelinum*	TM-1	DES119	HS46
G. ramondii							
3-79	21(7,0,1,1)[c]						
G. tomentosum	20(6,0,1,1)	2(2,0,0,0)					
G. mustelinum	17(4,1,1,1)	8(2,1,0,0)	8(2,1,0,0)				
TM-1	17(5,0,1,1)	6(1,0,0,0)	5(1,0,0,0)	6(1,1,0,0)			
DES119	17(5,0,1,1)	6(1,0,0,0)	5(1,0,0,0)	6(1,1,0,0)	0(0,0,0,0)		
HS46	17(5,0,1,1)	6(1,0,0,0)	5(1,0,0,0)	6(1,1,0,0)	0(0,0,0,0)	0(0,0,0,0)	
MARCABUCAG8US-1-88	17(5,0,1,1)	6(1,0,0,0)	5(1,0,0,0)	6(1,1,0,0)	0(0,0,0,0)	0(0,0,0,0)	0(0,0,0,0)

[a] The sequence length was 876 bp; 3′ untranslated region (3′-UTR) and exon length were 600 and 276 bp, respectively. In the exon, only one SNP sites had been found

[b] 3-79 is a double-haploid line of *G. barbadense* L.; TM-1 is a *G. hirsutum* L. inbred genetic standard line; DES119, HS46, and MARCABUCAG8US-1-88 are three *G. hirsutum* lines

[c] It denotes that 21 SNP sites had been found between *G. ramondii* and *G. barbadense* (3-79). Among them, seven were transversional SNP, no indel had been found between the two genotypes, one SNP located in the exon changed the amino acid

Table 2 SNP characters based on the *GhMyb10* gene fragment[a]

Genotype[b]	*G. arboreum*	3-79	*G. tomentosum*	*G. mustelinum*	TM-1	DES119	HS46
G. arboreum							
3-79	17(4,8,1,0)[c]						
G. tomentosum	17(4,8,1,0)	0(0,0,0,0)					
G. mustelinum	35(5,24,2,1)	18(1,16,1,1)	18(1,16,1,1)				
TM-1	20 (5,8,1,0)	3(1,0,0,0)	3(1,0,0,0)	21(2,16,1,1)			
DES119	20 (5,8,1,0)	3(1,0,0,0)	3(1,0,0,0)	21(2,16,1,1)	0(0,0,0,0)		
HS46	20 (5,8,1,0)	3(1,0,0,0)	3(1,0,0,0)	21(2,16,1,1)	0(0,0,0,0)	0(0,0,0,0)	
MARCABUCAG8US-1-88	26(5,14,1,0)	9(1,6,0,0)	9(1,6,0,0)	27(2,22,1,1)	6(0,6,0,0)	6(0,6,0,0)	6(0,6,0,0)

[a] The sequence length was 911 bp; 3′ untranslated region (3′-UTR) and exon length were 644 and 267 bp, respectively

[b] 3-79 is a double-haploid line of *G. barbadense* L.; TM-1 is a *G. hirsutum* L. inbred genetic standard line; DES119, HS46, and MARCABUCAG8US-1-88 are three *G. hirsutum* lines

[c] It denotes that 17 SNP sites had been found between *G. arboreum* and *G. barbadense* (3-79). Among them, four were transversional SNP and eight indel had been found between the two lines. Only one SNP is located in the exon, and the SNP doesn't change the coded amino acid

following locations: long arm of chromosome 14, long arm of chromosome 15, chromosome 19, 21, or 24. Similarly, *GhMyb10* could be located on long arm of chromosome 5, 8, 11 or chromosome 13. Further investigation is needed to confirm these putative assignments of the chromosomal locations.

Expression pattern of *GhMyb8/10* gene in cotton tissues

Northern blot analysis was used to investigate the expression pattern of *GhMyb8/10* gene. The nonconserved TRR region of *GhMyb8/10* was used as a hybridization probe to prevent cross hybridization with other cotton *Myb* genes in the Northern blot analysis. Northern analysis results (Fig. 4) showed that a 1.03-kb *GhMyb8/10* mRNA was detected in all cotton tissues, including leaves, flowers, roots, and fibers from different developmental stages (5, 10, 15, and 20 DPA); however, the expression levels were more abundant in flowers and roots.

Overexpression of *GhMyb10* caused morphological changes in the trichome cells of transgenic tobacco plants

The standard *Agrobacterium*-mediated leaf-disc transformation method was used to generate transgenic tobacco plants. The expression level of *GhMyb10* in transgenic plants was determined by Northern analysis using the gene-specific region (C-terminal TRR region) of *GhMyb10* as a hybridization probe. As a negative control, no hybridization signal was detected in the wild type tobacco plant, indicating that there was no *GhMyb10* homologous gene present in the tobacco genome (Fig. 5). In comparison with the wild type tobacco plants, the *Pro*$_{35S}$: *GhMyb10* transgenic plants showed normal pheno-

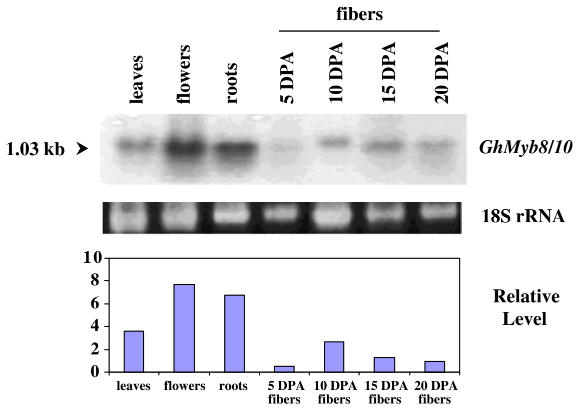


Fig. 4 Northern blot analysis of *GhMyb8/10*. Total RNA (10 μg) isolated from leaves, flowers, roots, and fibers at 5 DPA, 10 DPA, 15 DPA, and 20 DPA were electrophoresed on an agarose gel and hybridized with the TRR region of *GhMyb10*. The EtBr-stained RNA gel is included as a loading control. The relative transcript levels of *GhMyb8/10* were determined by the ratio of hybridized intensity of the 1.03-kb *GhMyb10* transcript to the EtBr-stained 18S rRNA band using Scion Image program (Scion Corporation, http://www.scio-corp.com)

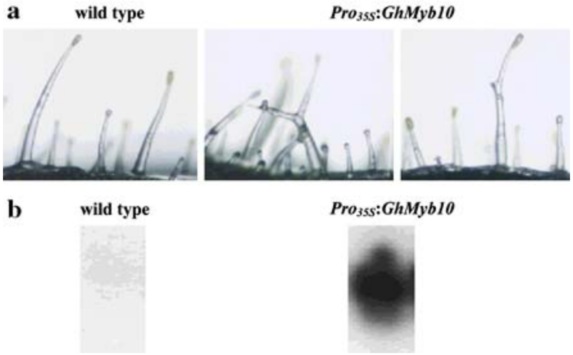


Fig. 5 Effect of *GhMyb10* overexpression on morphology of leaf trichomes in transgenic tobacco plants. The abnormal cell shapes (Y and irregular shapes) of leaf trichomes in transgenic lines that overexpress the *GhMyb10* gene were observed in comparing with the wild type tobacco plant (**a**). The corresponding *GhMyb10* transcript level of wild type and transgenic tobacco plants is analyzed by Northern blotting (**b**)

types in general and no effect on development processes (data not shown). However, the morphology of leaf trichomes from *Pro*$_{35S}$:*GhMyb10* transgenic plants was different from the wild type plant (Fig. 5). Some leaf trichomes from *Pro*$_{35S}$:*GhMyb10* transgenic plants showed a Y or irregular shape in addition to the single-spike shape normally found in wild type plants (Fig. 5). These results suggest that the overexpression of the *GhMyb10* gene altered the pathway of epidermal cell differentiation and caused abnormal cell shapes of leaf trichomes in tobacco plants.

Discussion

The unique plant *R2R3-MYB* gene family is one of the largest and most diverse super-gene families in the plant kingdom, and it is believed that this diverse super-gene family is mainly involved in the "plant-specific" processes (Martin and Paz-Ares 1997; Jin and Martin 1999). In this study, two cotton *R2R3-MYB* genes (*GhMyb8* and *GhMyb10*) were isolated and characterized. Our results of genomic DNA sequences (Fig. 1) and genomic origin analysis (Fig. 3b) indicated that *GhMyb8* and *GhMyb10* are alloallelic genes in the allotetraploid cotton. Several studies (Grula et al. 1995; Ferguson et al. 1997; Small and Wendel 2000) have suggested that the diploids *G. herbaceum* (A_1) and *G. raimondii* (D_5) are A and D genome donors of the tetraploid *G. hirsutum* (AD), respectively. However, our genomic origin analysis (Fig. 3b) showed that *GhMyb10* gene was present in A_2 (*G. arboreum*) and AD (*G. hirsutum*) genomes, whereas *GhMyb8* in D_5 (*G. raimondii*) and AD (*G. hirsutum*) genomes. The SNP analysis also showed sequence similarity between *G. raimondii* (D_5) and *G. hirsutum* (AD) species in *GhMyb8* and *G. arboreum* (A_2) and *G. hirsutum* (AD) species in *GhMyb10* (Tables 1 and 2). These results suggest that *G. raimondii* (D_5) and *G. arboreum* (A_2), and not the A_1 genome from *G. herbaceum*, might be the possible ancestral D genome and A genome donors for the allotetraploid *Gossypium hirsutum* (AD genome).

The *GhMyb8/10* transcript was present in all examined tissues, including leaves, flowers, roots, and fibers from different developmental stages in Northern blotting analysis (Fig. 4). Other cotton *R2R3-Myb* genes (*GhMyb1*, *2*, and *3*) have also been found to exhibit a global expression patter (Loguercio et al. 1999). However, the Northern blotting analysis couldn't distinguish the expression differences between these two alloallelic genes because of their sequence similarity. The cotton *GhMYB1* gene has recently reported to exhibit different expression levels between two homologous genes in the allopolyploid cotton (Cedroni et al. 2003). It will be interesting to examine whether this allelic-specific expression pattern also exists between *GhMyb8* and *GhMyb10*.

The ectopic expression of the *GhMyb10* gene affects the epidermal cell differentiation of trichome cells in transgenic tobacco plants (Fig. 5). Similarly, the cotton *R2R3-MYB* genes, *GhMyb1* (formerly called *CotMYBA*) (Loguercio et al. 1999; Payne et al. 1999) and *GhMyb25* (Wu et al. 2006), which are close homologs of *GhMyb10*, have also been reported to affect trichome differentiation when they were overexpressed in tobacco plants. Many transcription factors, including *MYB* proteins, bHLH (basic helix-loop-helix) factors, and WD-40 repeats, are involved in the processes of asymmetric cell division and intercellular signaling to regulate the patterning of different cell types in epidermal cells (Lee and Schiefelbein 1999; Serna and Martin 2006). In *Arabidopsis*, this MYB-bHLH-WD40 regulatory complex, including *GLABRA1* (*GL1*) (Oppenheimer et al. 1991), *TRANSPARENT TESTA GLABRA1* (*TTG1*) (Koornneef 1981), *GLABRA3* (*GL3*) (Koornneef et al. 1982; Payne et al. 2000), *TRIPTYCHON* (*TRY*) (Hulskamp et al. 1994), and *AtMYB23* (Kirik et al. 2001), controls the initiation and spacing of trichome cells. The fiber development belongs to one kind of epidermal cell differentiation. The regulation of fiber development by a similar network of transcription factors has been proposed (Serna and Martin 2006), and the *GhMyb10* gene might be part of this network. Further experiments and analyses will confirm whether *GhMYB10* has a functional role in cotton fiber differentiation and development.

The 125 R2R3 *Myb* members in *A. thaliana* have been classified as 22 subgroups based on the phylogenetic analysis of the first 320 N-terminal amino acid sequences of the *AtMYB* proteins (Kranz et al. 1998; Stracke et al. 2001). Both *GhMyb1* (Wilkins and Zhou 2002) and *GhMyb10* can be grouped into the subgroup 4 of the *A. thaliana R2R3-MYB* gene family as shown in Fig. 2 (Kranz et al. 1998; Stracke et al. 2001). Except for the conserved DNA-binding domain, all subgroup 4 members contain other conserved amino acid motifs, including LlsrGIDPxT/SHRxI/L, pdLNLD/ElxiG/S, and $CX_{1-2}CX_{7-12}CX_2C$ (Zn-finger) (upper case letters represent aa residues present in all members of the subgroup, lower case letters indicate aa residues conserved in more than 50% of the gene members, and X represents any aa residue), in the diverse C-terminal domain. The subgroup 4 *Myb* genes, including *AtMyb4* (Jin et al. 2000), *AmMyb308* (Tamagnone et al. 1998a, 1998b), *AmMyb330* (Tamagnone et al. 1998a, 1998b), *GhMyb1* (Wilkins and Zhou 2002), *ZmMyb31* and *ZmMyb43* (Fornalé et al. 2006), have been shown to regulate the biosynthesis of phenylpropanoids. These six genes act as negative regulators to repress the synthesis of hydoxycinnamic acid and lignin. The caffeic acid O-methyl-transferase gene (*COMT*) involved in lignin biosynthesis has been found to be down-regulated by *ZmMYB31* and *ZmMYB43* in transgenic maize and Arabidopsis plants (Fornalé et al. 2006). The functional domain analysis of the *AtMyb4* gene indicates that the small conserved motif (pdLNLD/ElxiG/S) might be responsible for the function of a repressor (Jin et al. 2000). *GhMYB10* shares similar transcript expression patterns with *AtMYB4* and *AtMYB32* and contains similar structural features in the C-terminal region as the six members of subgroup 4 *MYB* proteins, suggesting that *GhMyb8/10* may also play a role in the phenylpropanoid metabolism.

SNP were detected from the partial fragments of *GhMyb8 and GhMyb10* genes at a mean frequency of 3.75% (one SNP per 27 bp sequence) among eight cotton genotypes from five species. Most of the SNP were found at the interspecies level as expected. Only six out of 67 SNP were found among the four *G. hirsutum* lines suggesting limited variation of *GhMyb8* and *GhMyb10* in cultivated cotton and their critical role in fiber development. In addition, significant uneven SNP occurrence frequencies were detected between coding region (0.55%) and 3′-UTR (5.14%). The uneven distribution of SNP in coding and non-coding regions was also identified in other crops (Kanazin et al. 2002; Zhu et al. 2003; Salmaso et al. 2004). The lower SNP frequency in the coding region indicated the conserve character of these genes in different species. The varied SNP frequency among species and across different region of genome had been discovered in cotton and other plants. Rong et al. (2004) sequenced total of 5409 bp sequence-tagged sites (STSs) in four tetraploid cotton genotypes and found that the rate of variation per nucleotide was 0.35% between the *G. hirsutum* and *G. barbadense* species, and 0.14% and 0.37%, respectively, between genotypes within species. One SNP occurring in 70, 78, 189, and 9 bp had been observed in the particular gene(s) of maize (*Zea mays* L.), grapevine (*Vitis vinifera*), barley (*Hordeum vulgare*), and wheat (*Triticum aestivum*), respectively (Ching et al. 2002; Kanazin et al. 2002; Caldwell et al. 2004; Salmaso et al. 2004). The sequence variation among species was indicated in the phylogentic analysis results (data not shown), which were rooted by the two diploid ancestral species respectively. The topologies difference from the well-established organismal history indicated the independent and specific evolving pattern of the two genes in A and D genome in tetraploid cotton after the polyploid event (Cronn et al. 1999; Wendel and Cronn 2003). The lack of genetic diversity in cotton has hindered the construction of genome-wide linkage map. Using SNP markers derived from candidate genes associated with fiber development, like *GhMYB* transcription factors, in molecular mapping project would expedite the discovery on the association of candidate genes with fiber traits.

Acknowledgement We are grateful to Dr. John Yu for providing *G. arboreum* and *G. ramondii* DNA samples, Mr. Douglas A. Dollar for the technical assistance in SNP genotyping and Ms. Yufang Guo for the kind suggestions on data analysis. This article was approved for publication as No. J-11076 of Mississipi Agricultural and Forestry Experiment Station.

References

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91

Brookes AJ (1999) The essence of SNPs. Gene 234:177–186

Caldwell KS, Dvorak J, Lagudah ES, Akhunov E, Luo MC, Wolters P, Powell W (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. Genetics 167:941–947

Cedroni ML, Cronn RC, Adams KL, Wilkins TA, Wendel JF (2003) Evolution and expression of *MYB* genes in diploid and polyploid cotton. Plant Mol Biol 51:313–325

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Kanazin V, Talbert H, See D, DeCamp P, Nevo E, Blake T (2002) Discovery and assay of single-nucleotide polymorphisms in barley (*Hordeum vulgare*). Plant Mol Biol 48:529–537

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, Hubbell E, Au M, Chung EY, Lashkari D, Lemieux B, Dean C, Lipshutz RJ, Ausubel FM, Davis RW, Oefner PJ (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat Genet 23:203–207

Cone KC, Cocciolone SM, Burr FA, Burr B (1993) Maize anthocyanin regulatory gene *pl* is a duplicate of *c1* that functions in the plant. Plant Cell 5:1795–1805

Cronn R, Cedroni M, Haselkorn T, Grover C, Wendel JF (2002a) PCR-mediated recombination in amplification products derived from polyploid cotton. Theor Appl Genet 104:482–489

Cronn RC, Small RL, Haselkorn T, Wendel JF (2002b) Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am J Bot 89:707–725

Cronn RC, Small RL, Wendel JF (1999) Duplicated genes evolve independently after polyploid formation in cotton. Proc Natl Acad Sci USA 96:14406–14411

Feinberg A, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6–13

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies India and Japonica genome alignments. Genome Res 14:1812–1819

Ferguson DL, Turley RB, Kloth RH (1997) Identification of a δ-TIP cDNA clone and determination of related A and D genome subfamilies in *Gossypium* species. Plant Mol Biol 34:111–118

Fornalé S, Sonbol F-M, Maes T, Capellades M, Puigdomènech P, Rigau J, Caparrós-Ruiz D (2006) Down-regulation of the maize and *Arabidopsis thaliana* caffeic acid 0-methyl-transferase genes by two new maize R2R3-MYB transcription factors. Plant Mol Biol 62:809–823

Gallagher SR (1992) GUS protocols: using the GUS gene as a reporter of gene expression. Academic Press, Inc, San Diego, CA

Glover BJ, Perez-Rodriguez M, Martin C (1998) Development of several epidermal cell types can be specified by the same MYB-related plant transcription factor. Development 125:3497–3508

Gonda TJ (1998) The c-Myb oncoprotein. J Biochem & Cell Biol 30:547–551

Grula JW, Hudspeth RL, Hobbs SL, Anderson DM (1995) Organization, inheritance and expression of acetohydroxyacid synthase genes in the cotton allotetraploid *Gossypium hirsutum*. Plant Mol Biol 28:837–846

Hsu C-Y, Jenkins JN, Saha S, Ma D-P (2005) Transcriptional regulation of the lipid transfer protein gene *LTP3* in cotton fibers by a novel MYB protein. Plant Sci 168:167–181

Hughes DW, Galau G (1988) Preparation of RNA from cotton leaves and pollen. Plant Mol Biol Reporter 6:253–257

Hulskamp M, Misera S, Jurgens G (1994) Genetic dissection of trichome cell development in *Arabidopsis*. Cell 76:555–566

Jefferson RA, Kacanagh TA, Beva MW (1987) GUS fusion: β-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. EMBO J 6:3901–3907

Jin H, Cominelli E, Bailey P, Parr A, Mehrtens F, Jones J, Tonelli C, Weisshaar B, Martin C (2000) Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in *Arabidopsis*. EMBO J 19:6150–6161

Jin H, Martin C (1999) Multifunctionality and diversity within the plant *MYB*-gene family. Plant Mol Biol 41:577–585

Kanazin V, Talbert H, See D, DeCamp P, Nevo E, Blake T (2002) Discovery and assay of single-nucleotide polymorphisms in barley (*Hordeum vulgare*). Plant Mol Biol 48:529–537

Kim MY, Van K, Lestart P, Moon JK, Lee SH (2005) SNP identification and SNAP marker development for a *GmNARK* gene controlling supernodulation in soybean. Theor Appl Genet 110:1003–1010

Kirik V, Schnittger A, Radchuk V, Adler K, Hulskamp M, Baumlein H (2001) Ectopic expression of the *Arabidopsis AtMYB23* gene induces differentiation of trichome cells. Develop Biol 235:366–377

Koornneef M (1981) The complex syndrome of *ttg* mutants. Arabidopsis Inf Serv 18:45–51

Koornneef M, Dellaert LWM, van der Veen JH (1982) EMS- and radiation-induced mutation frequencies at individual loci in *Arabidopsis thaliana*. Mutat Res 93:109–123

Kranz HD, Denekamp M, Greco R, Jin H, Levya A, Meissner RC, Petroni K, Urzaingui A, Bevan M, Martin C, Smeekens S, Tonelli C, Paz-Ares J, Weisshaar B (1998) Towards functional characterization of the members of the *R2R3-MYB* gene family from *Arabidopsis thaliana*. Plant J 16:263–276

Kumar P, Paterson AH, Chee PW (2006) Predicting intron sites by aligning cotton ESTs with *Arabidopsis* genomic DNA. J Cotton Sci 10:29–38

Lee MM, Schiefelbein J (1999) WEREWOLF: a MYB-related protein in *Arabidopsis*, is a position-dependent regulator of epidermal cell patterning. Cell 99:473–483

Liu S, Saha S, Stelly DM, Burr B, Cantrell RG (2000) Chromosomal assignment of microsatellite loci in cotton. J Hered 91:326–332

Loguercio LL, Zhang J-Q, Wilkins TA (1999) Differential regulation of six novel *MYB*-domain genes defines two distinct expression patterns in allotetraploid cotton (*Gossypium hirsutum* L.). Mol Gen Genet 261:660–671

Martin C, Paz-Ares J (1997) MYB transcription factors in plants. Trends Genet 13:67–73

Noda K, Glover BJ, Linstead P, Martin C (1994) Flower colour intensity depends on specialized cell shape controlled by a *Myb*-related transcription factor. Nature 369:661–664

Oppenheimer OG, Hermn PL, Sivakumaran S, Esch J, Marks DM (1991) A *myb* gene required for leaf trichome differentiation in *Arabidopsis* is expressed in stipules. Cell 67:483–493

Paterson AH, Brubaker CL, Wendel JF (1993) A rapid method for extraction of cotton (*Gossypium spp.*) genomic DNA suitable for RFLP or PCR analysis. Plant Mol Biol Reporter 11:122–127

Payne T, Clement J, Arnold D, Lloyd A (1999) Heterologous myb genes distinct from *GL1* enhance trichome production when overexpressed in *Nicotiana tabacum*. Development 126:671–682

Payne CT, Zhang F, Lloyd AM (2000) *GL3* encodes a bHLH protein that regulates trichome development in *Arabidopsis* through interaction with GL1 and TTG1. Genetics 156:1349–1362

Paz-Ares J, Ghosal D, Wienand U, Peterson PA, Saedler H (1987) The regulatory *c1* locus of *Zea mays* encodes a protein with homology to *myb* proto-oncogene products and with structural similarities to transcriptional activators. EMBO J 6:3553–3558

Percival AE, Kohel RJ (1990) Distribution, collection, and evaluation of *Gossypium*. Adv Agron 44:225–256

Rafalski JA (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci 162:329–333

Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS, Park C, Pierce GJ, Rainey KM, Rastogi VK, Schulze SR, Trolinder NL, Wendel JF, Wilkins TA, Williams-Coplin TD, Wing RA, Wright RJ, Zhao X, Zhu L, Pateson AH (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). Genetics 166:389–417

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analysis by the coalescent and other methods. Bioinformatics 19:2496–2497

Salmaso M, Faes G, Segala C, Stefanini M, Salakhutdinov I, Zyprian E, Toepfer R, Grando MS, Velasco R (2004) Genome diversity and gene haplotypes in the grapevine (*Vitis vinifera*, L.), as revealed by single nucleotide polymorphisms. Mol Breed 14:385–395

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF (2003) Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Mol Biol Evol 20:633–643

Serna L, Martin C (2006) Trichomes: different regulatory networks lead to convergent structures. Trends Plant Sci 11:274–280

Shappley ZW, Jenkins JN, Meredith WR, McCarty JC (1998a) An RFLP linkage map of Upland cotton, *Gossypium hirsutum* L. Theor Appl Genet 97:756–761

Shappley ZW, Jenkins JN, Zhu J, McCarty JC (1998b) Quantitative trait loci associated with agronomic and fiber traits of Upland cotton. J Cotton Sci 2:153–163

Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA (1995) An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res 23:1087–1088

Small RL, Wendel JF (2000) Copy number liability and evolutionary dynamics of the *Adh* gene family in diploid and tetraploid cotton (*Gossypium*). Genetics 155:1913–1926

Stracke R, Werber M, Weisshaar B (2001) The *R2R3-MYB* gene family in *Arabidopsis thaliana*. Curr Opin in Plant Biol 4:447–456

Suo J, Liang X, Pu L, Zhang Y, Xue Y (2003) Identification of *GhMYB109* encoding a R2R3 MYB transcription factor that expressed specifically in fiber initials and elongating fibers of cotton (*Gossypium hirsutum* L.). Biochim Biophys Acta 1630:25–34

Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts

Tamagnone L, Merida A, Parr A, Mackay S, Culianez-Macia FA, Roberts K, Martin C (1998a) The AmMYB308 and AmMYB330 transcription factors from *Antirrhinum* regulate phenylpropanoid and lignin biosynthesis in transgenic tobacco. Plant Cell 10:135–154

Tamagnone L, Merida A, Stacey N, Plaskitt K, Parr A, Chang C-F, Lynn D, Dow M, Roberts K, Martin C (1998b) Inhibition of phenolic acid metabolism results in precocious cell death and altered cell morphology in leaves of transgenic tobacco plants. Plant Cell 10:1801–1816

Ulloa M, Saha S, Jenkins JN, Meredith WR, McCarty JC, Stelly DM (2005) Chromosomal assignment of RFLP linkage groups harboring important QTLs on an intraspecific cotton (*Gossypium hirsutum* L.) joinmap. J Hered 96:132–144

Van K, Hwang EY, Kim MY, Park HJ, Lee SH, Cregan PB (2005) Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. J Hered 96:529–535

Walkerpeach CR, Velten J (1994) *Agrobacterium*-mediated gene transfer to plant cells: Cointegrate and binary vector systems. In: Gelvin SB, Schilperoort RA (eds) Plant molecular biology manual. Kluwer Academic Publishers, Belgium, pp. B1/1–B1/19

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Wang S, Wang J-W, Yu N, Li C-H, Luo B, Gou J-Y, Wang L-J, Chen X-Y (2004) Control of plant trichome development by a cotton fiber MYB gene. Plant Cell 16:2323–2334

Wendel JF, Cronn C (2003) Polyploidy and the evolutionary history of cotton. Adv Agron 78:139–186

Wilkins TA, Rajasekaran K, Anderson DM (2000) Cotton biotechnology. Crit Rev Plant Sci 19:511–550

Wilkins TA, Zhou F (2002) A novel cotton fiber R2R3-MYB transcription factor represses phenylpropanoid biosynthesis via a unique genetic mechanism, In: Plant Biology Conference, American Society of Plant Biologists

Wu Y, Machado AC, White RG, Llewellyn DJ, Dennis ES (2006) Expression profiling identifies genes expressed early during lint fiber initiation in cotton. Plant Cell Physiol 47:107–127

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134

Zimdahl H, Nyakatura G, Brandt P, Schulz H, Hummel O, Fartmann B, Brett D, Droege M, Monti J, Lee YA, Sun Y, Zhao S, Winter EE, Ponting CP, Chen Y, Kasprzyk A, Birney E, Ganten D, Hubner N (2004) A SNP map of the rat genome generated from cDNA sequences. Science 303:807